ESERCITAZIONE 3 – TEXT SUMMARIZATION

INTRODUZIONE

La terza esercitazione prevede la realizzazione del task di summarization. Il task consiste nell'effettuare un riassunto automatico a partire da un testo in input. La risorsa utilizzata per svolgere il riassunto è Nasari in formato embedded, ovvero una rappresentazione vettoriale di synset Babelnet.

L'approccio utilizzato è detto **estrattivo** e di tipo statistico: il riassunto viene creato estraendo parti di testo rilevanti (interi paragrafi o frasi) in base al **tasso di compressione** utilizzato. Il tasso di compressione indica la quantità di testo che deve essere ignorata per far spazio ad informazioni rilevanti. Sono stati utilizzati tassi di compressione del 10,20 e 30%.

Il riassunto deve essere **indicativo**, ossia deve fornire un'idea sul contenuto del documento; **informativo**, cioè deve contenere tutti gli elementi principali e rilevanti; **critico**, ossia valutare il documento esprimendo una visione sulla qualità del lavoro dell'autore.

SVILUPPO

I criteri di rilevanza sono fondamentali per la creazione del topic. Dopo una prima fase di analisi dei possibili criteri, sono stati utilizzati due principali criteri:

- **Metodo del titolo (TITLE)**
  l titolo del documento dà informazioni sul contenuto. È possibile utilizzare le parole del titolo (eliminando le stop-words) per trovare il contenuto importante all'interno del testo e utilizzare queste parole chiave per le ricerca di paragrafi importanti nel documento. Tuttavia, è sbagliato sperimentare esclusivamente tramite questo metodo perchè il titolo in alcuni testi potrebbe non essere disponibile.
- **Cue phrases method (CUE)**
  Nei testi ci sono delle frasi che contengono parole che permetto di capire che stanno per essere dette cose importanti (bonus phrases) o inutili (stigma phrases). A tal proposito sono state individuate due liste sulla base di una ricerca: una contenente **bonus word** (composta prevalentemente da comparativi, superlativi, avverbi...) e l'altra contenente **stigma word** (composta esclusivamente da pronomi). Queste frasi possono essere trovare automaticamente assegnando un punteggio ad una frase. Il punteggio di una sentence aumenta, se contiene bonus phrase; diminuisce se contiene stigma phrase. Questi dati sono stati utilizzati per indurre un ranking dei paragrafi del documento da riassumere, sulla base della presenza di queste parole, utile al fine di individuare il topic del documento come il paragrafo classificato in prima posizione.

L'algoritmo summarization è diviso in due fasi:

- **Creazione del topic semantico**
  La creazione del topic consiste nell'utilizzare uno dei due criteri di rilevanza per ottenere una bag of words di termini rilevanti. Una volta ottenuta, vengono associate **informazioni semantiche** sfruttando il file dd-small-nasari-15.txt, attraverso il quale è possibile ottenere un dizionario associando ad ogni termine rilevante un insieme di **vettori Nasari** che rappresentano ognuno un **synset** di **BabelNet**; l'associazione viene effettuata facendo matchare ogni termine del topic con il WikiTitle del vettore o dei vettori Nasari corrispondente/i. La funzione *get_topic* o *get_title_topic* effettua l'operazione appena descritta.

- **Ranking dei paragrafi e Weighted Overlap**
  I vettori Nasari utilizzati sono **vettori lessicali**: la descrizione vettoriale è realizzata mediante un numero di termini (**features**) associati ad un peso. Il peso induce un ranking di importanza delle features al suo interno.

  Dopo aver calcolato il topic semantico, attraverso la funzione *get_context_paragraph* viene calcolato il **contesto semantico** per ogni paragrafo, ottenendo un dizionario analogo a quello del topic in cui ad ogni termine (chiave) corrisponde un insieme di vettori NASARI (valore).

  A questo punto è necessario calcolare uno **score** per ogni paragrafo. Lo score è calcolato sommando iterativamente la **similarità tra due concetti**, uno individuato dalla parola nel contesto semantico del paragrafo e l'altro individuato dalla parola nel topic: in pratica calcola la similarità tra ogni parola del topic e ogni parola del contesto del paragrafo, dove ogni parola è espressa da un insieme di vettori Nasari. La sommatoria viene divisa per il numero di confronti effettuati ottenendo così uno score medio da associare direttamente al paragrafo.

  La funzione *similarity* calcola la similarità tra due termini massimizzando il **Weighted Overlap** tra i vettori associati, maggiore è l'overlap è più simili saranno i vettori. Ad esempio, dati Cw1 = {c1, ..., cn} e Cw2 = {c1, ..., cm} insiemi di concetti (vettori NASARI) associati rispettivamente ai termini w1 e w2 è possibile calcolare la similarità:

$$sim(w_1, w_2) = \max_{v_1 \in \mathcal{C}_{w_1}, v_2 \in \mathcal{C}_{w_2}} \sqrt{WO(v_1, v_2)}$$

  dove:

$$WO(v_1, v_2) = \frac{\sum_{q \in O}(rank(q, v_1) + rank(q, v_2))^{-1}}{\sum_{i=1}^{|O|}(2i)^{-1}}$$

  abbiamo **O** come insieme di features comuni ai due vettori. Se le features comuni sono ugualmente importanti, cioè hanno una posizione (**rank**) simile nel vettore, allora il WO sarà più alto.

VALUTAZIONE DEI RISULTATI

La valutazione del riassunto è stata effettuata seguendo tre passaggi:

- Ricerca delle parole più importanti all'interno del documento originale utilizzando lo score **tf-idf** per stilare una classifica e scegliere la percentuale dei termini in base al tasso di compressione utilizzato. Ad esempio, se si utilizza il tasso di compressione del 20% e la classifica dei termini con il loro tf-idf ha cardinalità 100, verranno scelte le prime 80 parole;
- Vengono ricavate tutte le parole del riassunto generato dal sistema;
- Vengono calcolate **Precision** e **Recall** sulla base dei due insiemi di parole appena calcolati, utilizzando due metriche distinte: **Bleu** e **Rouge.**

Il BLEU (Precision) originalmente veniva utilizzato per misurare la qualità di una traduzione fatta da una macchina confrontandola con quello che avrebbe fatto un essere umano esperto. Calcola quanti 1-gram del riassunto sono presenti anche nel documento originale.

La metrica del Rouge (Recall) si basa sullo stesso principio del BLEU, ovvero calcola quanto la "traduzione" della macchina sia simile alla traduzione da parte di un agente umano. A differenza di BLEU, calcola quanti 1-gram del documento originale sono presenti anche nel riassunto. La Recall, quindi, risulta essere il dato più indicativo.

L'algoritmo di **summmarization** è stato applicato su cinque file (Andy-Warhol.txt, Ebola-virus-disease.txt, Life-indoors.txt, Napoleon-wiki.txt, Trump-wall.txt) che trattano argomenti differenti. Di seguito vengono mostrati i risultati del documento Trump-wall.txt analizzato con entrambi i metodi realizzati.

Metodo CUE:

- **10%:** Bleu (precision): 0.89, Rouge (Recall): 0.91
- **20%:** Bleu (precision): 0.81, Rouge (Recall): 0.80
- **30%** Bleu (precision): 0.72, Rouge (Recall): 0.68

Metodo TITLE:

- **10%:** Bleu (precision): 0.89, Rouge (Recall): 0.91
- **20%:** Bleu (precision): 0.80, Rouge (Recall): 0.75
- **30%:** Bleu (precision): 0.65, Rouge (Recall): 0.59

Come prevedibile, man mano aumentiamo il tasso di compressione entrambe le metriche subiscono una riduzione più o meno significativa.

ROUGE è un'ottima metrica di valutazione, ma presenta alcuni inconvenienti. In particolare, ROUGE non si rivolge a parole diverse che hanno lo stesso significato, in quanto misura le corrispondenze sintattiche piuttosto che la semantica. Quindi, se avessimo due sequenze che avevano lo stesso significato, ma usavano parole diverse per esprimere quel significato, potrebbe essere assegnato loro un punteggio ROUGE basso.